Analysis of the Capital Structure of German Companies with the SSA and SVM

Master of Science Thesis in Statistics

presented by: Tim-Fabien Pohlmann
Matrikel-Nr. 190049
Email: timpohlmann@gmx.ch

First Advisor: Prof. Dr. Wolfgang Härdle
Second Advisor: Dr. Bernd Droge

handed in: June 11, 2007

Acknowledgement

The author gratefully acknowledges that this work was supported by the Deutsche Forschungsgemeinschaft (DFG) through the SFB 649 “Economic Risk”. The provision of the database used in this thesis has been kindly taken over by the Creditreform AG.

The author is grateful to R.A. Moro (C.A.S.E. - Center for Applied Statistics and Economics) and Dr. D. Belomestny (WIAS - Weierstrass Institute for Applied Analysis and Stochastics) for research supervision and offering valuable comments on earlier versions of the thesis.

Abstract

This thesis presents and compares the performance of two recently developed classification methods namely the Spatial Stagewise Aggregation procedure and Support Vector Machines. Both techniques are convenient for the application to corporate bankruptcy analysis, in terms of calculation of default probabilities. Repeated random selection simulations varying with respect to variable and record choices for both methods proved a clear superiority in terms of the hit rate, representing the percentage of correctly classified observations, in favor of the SVM. Moreover the thesis presents a way to derive recommendations with respect to the capital structure policy of German manufacturing industry firms on the basis of the evaluated default probabilities.

Contents

List of Figures

List of Tables

List of Abbreviations

AWS	Adaptive Weighted Smoothing
cf.	confer
cont.	continued
D	Debt of a Firm
EAD	Exposure At Default
EBT	Earnings Before Taxes
e.g.	for example
FEDC	Financial and Economic Data Center
HR	Hit Rate
i.e.	that means
IC	Costs of Insolvency
LCP	Local Change Point
LGD	Loss Given Default
LPA	Local Parametric Assumption
MM	Modigliani-Miller
PD	Probability of Default
resp.	respectively
SAS	Statistical Analysis System
SMB	Small Modeling Bias
SSA	Spatial Stagewise Aggregation
SVM	Support Vector Machines
TA	Total Assets
TS	Tax Shield
V	Firm Value
vs.	versus
WACC	Weighted Average Cost of Capital

List of Symbols

χ	Euclidean space
d	Dimension
$d(x, y)$	Distance measure
$\mathbb{E}$	Expectation operator
$E_k(\mathfrak{z})$	Confidence set
$f(x)$	Regression function
$f(X_i) \approx \theta$	Local parametric assumption
$f(x) \equiv f_\theta(x)$	Global Parametric Assumption
γ	Weighting parameter
Γ	Gamma function
h_i	Radius of the window U
$1, \ldots, k, \ldots, K$	Number of radii
$\mathcal{K}(P, Q)$	Kullback-Leibler-Divergence
$\mathcal{K}(\theta, \widetilde{\theta})$	Kullback-Leibler-Divergence between P_θ and $P_{\widetilde{\theta}}$
$K(x)$	Kernel function
K_{tr}	Triangle Kernel
$L(W, \theta)$	Likelihood function
$\widehat{m}_k(x)$	"Smooth" regression function
$\lvert\omega\rvert$	Euclidean norm
$p(\cdot, f(x))$	Density function
P	Probability measure of a single observations
P_θ	Parametric probability measure of a single observations
$\mathbb{P}_\theta$	Parametric probability measure of n observations
ρ	Confidence level
R_E	Cost of Equity
R_D	Cost of Debt
T	Tax

List of Symbols (cont.)

θ	Maximum likelihood estimator
$\widetilde{\theta}$	"Weak" estimate
$\widehat{\theta}$	Aggregated SSA estimate
$\mathfrak{r}$	Oracle risk
$\mathcal{S}$	Matrix of variances of all components
T_C	Corporate tax rate
U	Window
$U(x)$	Window of the design space depending on x
V_L	Value of the levered firm
V_U	Value of the unlevered firm
$W = \{w_1, \ldots, w_n\}$	Set of weights
X_i	Vector of explanatory variables
Y_i	Vector of response variables
ξ	SVM penalty term
$\mathfrak{z}_k$	Critical values

1 Introduction

A recently developed nonparametric classification method called *Spatial Stagewise Aggregation* (SSA) will be presented in this thesis. SSA goes back to Belomestny and Spokoiny (2006), who were initially inspired by the local likelihood approach presented by Fan, Farmen & Gijbels (1998). The idea of SSA is a further development of the oracle and kernel methods where the kernel shape is chosen adaptively. The delivered classification estimates from the SSA procedure are based on sequences of maximum likelihood (ML) estimators, which are integrated in an aggregation procedure generating so-called *ensemble* or *oracle* estimates with remarkable properties in terms of variability reduction. Furthermore, the field of ensemble estimators has been studied in a seminal way by Breiman (1996). Scientifically the contribution of the SSA is the pointwise (spatial) aggregation procedure.

Moreover, from a practical point of view classification methods tend to become more and more important in the field of corporate bankruptcy analysis and company rating due to the implementation of the new Basel II capital accord. In this context one of the purposes of the thesis at hand is to measure and to compare the performance of SSA with another non-parametric non-linear classification method, namely the Support Vector Machines (SVM).

Additionally, a principal issue discussed in this work besides the performance presentation of statistical classification methods will be the following: what is the optimal ratio between external and internal financing? According to Damodaran (2002) there are two possibilities: on one hand *debt financing* can be chosen which is usually less expensive than internal financing but involves fixed debt payments in the future. In this case the firm may go bankrupt if it fails to pay interest on its debt. On the other hand it has the opportunity to finance itself *internally*. In this case the firm can retain whatever cash flows are left after debt payments have

been made. Now, the following central question arises, is there the optimal mix of debt and equity, or, more precisely, is there the optimal capital structure, and how can it be attained? The next chapter introduces a possible way of determining this optimal capital structure by applying economic theory.

Overall, the purpose of the present work is threefold: *first*, it introduces a recently developed classification method named the Spatial Stagewise Aggregation (SSA) to calculate default probabilities (PD) for German firms of the manufacturing industries. At the core of the thesis is the parallel development and implementation of an algorithm in the Statistical Analysis System (SAS) for PD evaluation. *Second*, these results will be compared with those obtained with another classification method, the Support Vector Machine (SVM). By comparing both performances with respect to classification accuracy it is possible to detect the weak and strong points of both approaches. *Third*, it is possible to use the resulting default probabilities to derive assessments of the optimal capital structure, i.e. the ideal mix of debt and equity.

The thesis is divided in five sections. After the introduction the *second section* discusses the financial motivation (i.e. capital structuring) for applying classification methods. The *third section* introduces theoretical issues: it presents the SSA method and then the machine learning approach as the foundation of the SVM. The *fourth section* gives a description of the *Creditreform* data, the variables and the preprocessing applied. The *fifth section* outlines empirical results. *Lastly* the research work done will be summed up and concluded.

2 Business Financing

To operate successfully in a market a firm is required to strengthen its position by increasing sales and capitalization. The management also has to analyze carefully market developments and company financial structure in order to guard against insolvency. The study of the interaction between these two objectives, (i) *increasing capitalization and profitability* and (ii) the *prevention of bankruptcy* is valuable for the financial theory since it sheds light on the internal workings of a company.

Profitability is the most important feature of a successful business. However, to obtain a credit for financing the business profits should display a certain degree of stability. A formal rating procedure for assessing the credibility of a company becomes even more important within the framework of Basel II. The rating score indicates how probable it is that a firm becomes bankrupt. Depending on the rating, a credit institution can formulate the conditions for issuing a credit. In this perspective it becomes increasingly important for a firm to obtain a higher rating and to be aware of the steps it needs to take to improve it.

In principle, a firm becomes bankrupt when the value of its assets drops below the value of its debt, or when a firm is unable to service its debt (cf. Ross et al. (2003)). In this context financial theory introduces the *capital structure* of a firm as an important value representing the relationship between the firms debt and equity. A firm should be cautious when it seeks to increase its debt. Damodaran (2002) mentions two aspects: on one hand a firm can benefit from debt e.g. with tax shields allowed by the government. On the other hand debt may have a disincentive effect as it raises the possibility of bankruptcy, raises agency costs and, moreover, decreases future flexibility.

2.1 Capital Structure Policy

The first proposition of the Modigliani-Miller (MM) theorem in a world with taxes yields two important statements for understanding the business financing mechanics. The *first statement* says that the value of an unlevered firm V_U plus the interest tax shield, defined as the product of the corporate tax rate τ and the effective debt D of the firm, i.e. $(\tau \cdot D)$, is equal to the value of the levered firm V_L,

$$V_L \quad = \quad V_U + \tau \cdot D. \tag{2.1}$$

Going one step further this statement says that in a world with corporate taxes the capital structure of a firm *matters*. This stands in contrast to the MM Theorem in a world without taxes which suggests that the capital structure has no effect on the financing strategy of a firm.

The *second statement* of the first proposition deals with the weighted average cost of capital (WACC) and says that as a firm relies more heavily on debt the WACC is decreasing. This phenomenon becomes obvious as the formula is considered:

$$\text{WACC} = \left(\frac{E}{D+E}\right) \cdot R_E + \left(\frac{D}{D+E}\right) \cdot R_D \cdot (1-\tau). \tag{2.2}$$

Here R_E (resp. R_D) denotes the cost of equity (resp. debt) of a firm. The firm value V equals $(D+E)$, where E is the equity of a firm.

Immediate conclusions of these remarks are the following: debt financing turns out to inhere benefits in terms of a tax shield, that increases the value V_L of the levered firm. Moreover, the fact that debt becomes a part of the business strategy is likely to increase the discipline of management, since with higher debt only most

promising investments are realized. Figure 2.1 illustrates the relationship between

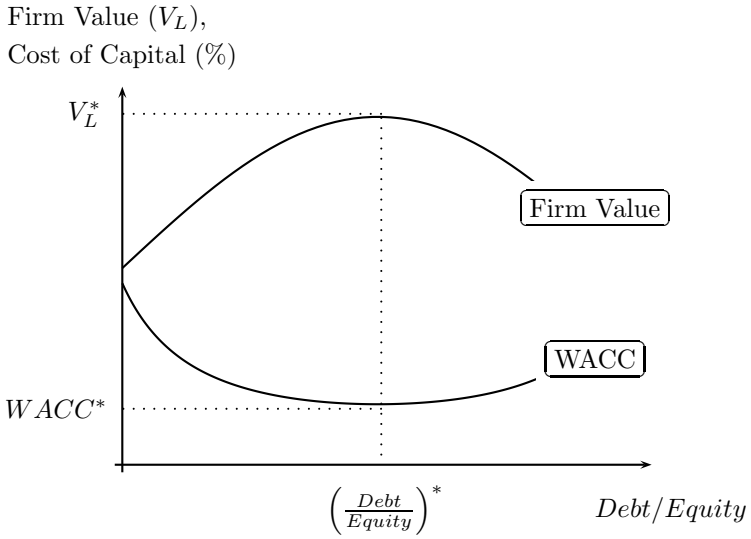

Fig. 2.1: Minimization problem

the firm value and WACC. The optimal value V_L of a levered firm is reached when the WACC is minimized. Upon closer inspection of figure 2.1 it is clear that higher leverage becomes disadvantageous as the firm value again decreases and capital costs rise. The disincentives originate from costs of financial distress (see Damodaran (2002)) and are composed of two components: *direct costs of bankruptcy* that emerge when a firm is liquidated and *indirect costs of bankruptcy* like agency costs of avoiding bankruptcy and restructuring a company (as under Chapter 11 of the U.S. Bankruptcy Code).

It is obvious that with increasing capitalization and probability of default (PD) *expected insolvency costs* (E[IC]) increase. As the debt grows financial benefits from tax shields will be counterbalanced by increasing costs of financial distress.

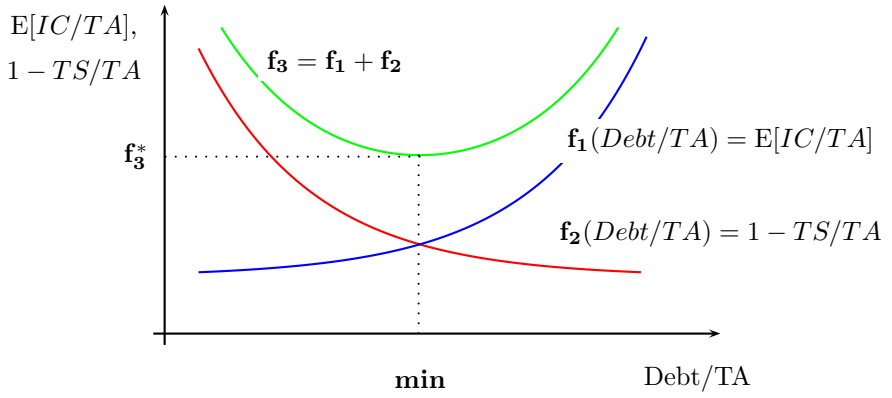

Fig. 2.2: Minimization problem: E$[IC/TA]$ denote the expected bankruptcy costs as the ratio of costs of insolvency (IC) and total assets (TA). TS/TA is the tax shield relative to assets. The functions $\mathbf{f_1}$ and $\mathbf{f_2}$ are examples.

Figure 2.2 illustrates the empirical relationship of the aforementioned expected bankruptcy costs and the tax shield. It shows that it is possible to detect the optimal firm value as in figure 2.1 with minimal costs of insolvency, by estimating the functions $\mathbf{f_1}$ and $\mathbf{f_2}$ that will be done in section 5. The only value that has to be calculated separately is the PD needed to compute expected bankruptcy costs.

2.2 Technical Aspects

The minimum capital costs as illustrated in figure 2.1 will be evaluated for the companies of the *Creditreform* dataset available through the *Financial and Economic Data Center* (FEDC). As no information is available concerning the tax T and tax shield TS it is necessary to define proxies for these values in order to estimate $\mathbf{f_2}$.

This can be done using the following approximate accounting identities

$$\begin{aligned} T_L &= \text{EBT} - \text{NI} \\ \tau &= \frac{T_L}{\text{EBT}} \end{aligned} \tag{2.3}$$

where EBT denotes earnings before taxes, and NI the net income of a firm. T_L is the estimated taxes payed by a levered firm, whereas (2.3) is the tax rate. The tax payed by an unlevered firm is

$$T_U = \tau \cdot \text{EBIT}.$$

Therefore it is possible to calculate the tax shield (TS) approximately as the difference:

$$\text{TS} = (T_U - T_L)^+.$$

The expected insolvency costs relative to total assets illustrated in figure 2.2 for the estimation of function $\mathbf{f_1}$ can be derived as the product of the probability of default (PD) and the loss given default (LGD), i.e.

$$\text{E}[IC/TA] = \text{PD} \cdot \text{LGD}. \tag{2.4}$$

3 Classification Techniques

The application of classification methods can be found in many, often independent of the contents, research fields. Unknowingly all of us come across them in everyday life. For instance, when using the internet, they help recognize web pages with the desired contents or filtrate spam. Very popular classification techniques are the generalized linear models (GLM), i.e. the logit model and some nonparametric regression and classification methods, i.e. the k-nearest neighbors (k-NN) method.

Both methods constituted the basis for further development of classification techniques. In recent years research in this field yielded approaches called *ensemble estimators* or *aggregated classification methods*. The name indicates the general idea of these methods when several individual estimators produce together ("*ensemble*") the final result. Most important in this category is Bagging (from Bootstrap Aggregating) which was developed by Breiman (1996) and (1998). Another popular technique in this context is called Boosting, and was first introduced by Freund (1995). Boosting is a so called voting method from computer science, more precisely from the field of machine learning. Generally speaking both methods are based on resampling techniques to obtain different training sets for each of the classifiers.

Depending on the context, conceivable alternatives to the aforementioned approaches can be the Spatial Stagewise Aggregation (SSA) and Support Vector Machines (SVM). Both alternatives will be treated in the two upcoming subsections 3.1 and 3.2. SSA can be placed very close to the k-NN method. The estimation procedure treats the so called "weak" estimates $\widetilde{\theta}$ based on a certain number of observations lying in the vicinity of a point x similarly to the k-NN method. After a detailed presentation of SSA and a brief description of SVM the performance of both methods will be compared by their application to the analysis of an extensive data set of German companies in the next chapter.

3.1 Spatial Stagewise Aggregation

Spatial Stagewise Aggregation (SSA) is a very recent classification and regression method which first was presented by Belomestny and Spokoiny (2006). It is based on modern learning theory. Conceptually it is close to another approach called Adaptive Weighted Smoothing (AWS) developed by Polzehl and Spokoiny (2000) which has been suggested in the context of image denoising where it enhances the quality of taken pictures with respect to contours. Their similarity lies in the "*propagation separation*" method as it was accurately called in Polzehl and Spokoiny (2005). Following this approach when trying to reveal an unknown local structure iteratively, the design allows to *propagate* in homogenous regions where local approximations are similar and *separates* between significantly different regions. Another appropriate designation of this procedure is the local change point (LCP) algorithm.

The SSA approach is also referred to as a *stagewise* aggregation algorithm. Stagewise in this context means that the estimates obtained at an earlier step are aggregated with the estimate from a new stage without changing the previous ones. An important property of the final estimates $\widehat{\theta}$ is that their pointwise risk does not exceed the smallest pointwise risk among all "weak" estimates up to a logarithmic multiple (Belomestny and Spokoiny, 2006). This property allows to obtain "optimal" aggregated estimates $\widehat{\theta}(x)$ in terms of fulfilling some kind of an "oracle" inequality.

The presentation is divided into two parts. First, the theoretical foundations of the SSA method will be presented in detail. The second part of this chapter giving a summary description and a presentation of the modular structure of the algorithm has a particular relevance for practical implementation.

3.1.1 Spatial Stagewise Aggregation as a Modern Nonparametric Classification Method

As it was mentioned above, the goal of the method is to detect homogenous regions of local approximations and separation patterns that help to distinct them from one another. According to this SSA suggests an approach of two different stages: the *first* stage of local adaptation, where an accurate localizing scheme in terms of *weights* is selected. At this stage a "weak" estimate $\widetilde{\theta}(x)$ is calculated, which depends locally on the set of points $\mathcal{X}$.
At the *second* stage the stepwise aggregation is done. During this procedure "weak" estimates are evaluated with respect to their homogenous character via the test statistic and the proper set of critical values.

3.1.2 Local Adaptation

To explain the local adaptation process it is indispensable to present briefly the underlying framework of local constant likelihood estimation. Let $Z_i = (Y_i, X_i)$ with $i = 1, \ldots, n$ be two vectors of random variables, where X_i's are explanatory variables or locations, valued in the finite Euclidean space $\mathcal{X} = I\!R^D$. Here the X_i's determine the distribution of the corresponding vector of observations Y_i. Now, a regression-like model can be formulated, where the Y_i's can be seen as responses, conditioned on $X_i = x$, and a density $p(\cdot, f(x))$. The following model can be stated as

$$Y_i \sim I\!P_{f(X_i)} \tag{3.1}$$

varying with values of $X_i = x$. The varying behavior of Y_i in the sense of its distribution, can be described by the parameter θ depending on X_i. The purpose

here is to make inferences using the "regression" function $f(x)$ (cf. Belomestny and Spokoiny (2006)). In this context the desired parameter θ, which by assumption determines the distribution of the observed responses Y_i should be estimated locally in a nonparametric way for a simple reason: if it were estimated globally with a classical ML approach, like

$$\widetilde{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log \mathbb{P}(Y_i, \theta) \tag{3.2}$$

it would yield a parameter $\widetilde{\theta}$ that would be a maximum over the whole range of explanatory variables, and would fail to detect homogeneity *within* different regions of the values of X. An alternative local parametric approach turns out to be more appropriate for the underlying purposes in this case: this would be a local adaptation to a constant value of the parameter $\widetilde{\theta}$ *within a specific neighborhood* of the point x. For its application it is necessary to assume the "smoothness property" of the regression function, which means that it changes only little in the vicinity of a point x (for a detailed treatise of this topic see Härdle (1990)). It should be mentioned at this point that a local constant design was chosen instead of a possible local linear design. As it was pointed out in Belomestny and Spokoiny (2006) the latter alternative revealed a decline with respect to estimation results.

3.1.3 Localization by Weights

As its name already suggests, the SSA method relies fundamentally, just as in nonparametric theory, on the idea of a *weighted* smoothing mechanism. Weights w_i are used in the model as it is considered to be the most general way to describe a model locally. In the present framework $f(X_i) \approx \theta$ being valid only approximately and in a small neighborhood of each point x (Local Parametric Assumption, LPA) as it would be too restrictive to assume $f(x) \equiv f_\theta(x)$ (Global Parametric Assumption)

(Härdle and Spokoiny, 2007). Thus it is possible to reformulate equation (3.2)

$$\widetilde{f}(x) = \underset{\theta\in\Theta}{\operatorname{argmax}} \underbrace{\sum_{i=1}^{n} w_i(x) \log \mathbb{P}_\theta(Y_i)}_{=L(W,\theta)} . \tag{3.3}$$

A striking intuition in terms of form and functioning of equation (3.3) can be obtained by considering the general formulation of a k-NN estimator

$$\widehat{m}_k(x) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} w_i(x) Y_i \quad \text{with} \quad w_i(x) = \mathbf{1}(x_i \in U_k(x)). \tag{3.4}$$

This is an estimator of the "smooth" regression function $\widehat{m}_k(x)$ and can be interpreted as a weighted average over the k *Nearest Neighbors* (k-NN) of a point. The corresponding k-neighborhood contains a certain quantity of points, at which a distance function, in general the Euclidean L_2-norm, measures the proximity of all points to each other. A decisive role in both equations (3.4) as well as in (3.3) is attached to the set of weights $W = \{w_1, \ldots, w_n\}$, since the choice of an appropriate weighting scheme exerts a major influence on the quality of the derived estimator. For the local weighted ML estimation of the function $\widetilde{f}(x)$ in equation (3.3) every observation is assigned to a weight $w_i(x)$.

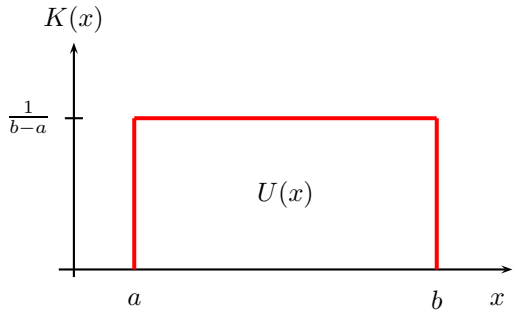

Fig. 3.1: Uniform kernel $K(x) = \frac{1}{b-a}\mathbf{1}(X_i \in U(x))$

The applied weighting scheme here reminds again of the k-NN estimator. How-

ever, the weights come here by choosing a window $U = U(x)$ of the design space depending on x, so that weights of the form

$$\begin{aligned} w_i(x) &= \frac{1}{|U(x)|}\mathbf{1}(X_i \in U(x)) \\ \text{with} \quad \sum_{i=1}^{n} w_i(x) &= 1 \end{aligned}$$

are obtained, observations lying outside of this window are not considered. This specific form of a weighting scheme goes back to the concept of kernel estimators, more precisely in the present case to the uniform kernel estimator (cf. Fig. 3.1). The particular choice of this kernel was done as it is a very convenient one in terms of algorithmic implementation. Moreover Belomestny and Spokoiny (2006) stress the fact that the choice of the kernel only has a minor effect on the estimation results.

3.1.4 Calculation of the Local Likelihood Estimates $\widetilde{\theta}$

The present case highlights the binary response (Bernoulli) model, where the responses Y_i are an i.i.d. Bernoulli random vector satisfying $\mathbb{P}_\theta(Y_i = 1) = \theta$ and $\mathbb{P}_\theta(Y_i = 0) = 1 - \theta$. The solution of the local likelihood problem stated by equation (3.3) can be obtained by

$$\begin{aligned} \underset{\theta\in\Theta}{\operatorname{argmax}}\, L(W,\theta) &= \log \prod_{i=1}^{n} \theta^{w_i Y_i}(1-\theta)^{w_i(1-Y_i)} \\ &= \log\theta \sum_{i}^{n} w_i Y_i + \log\,(1-\theta) \sum_{i}^{n} w_i(1-Y_i) \\ &= \underbrace{\sum_{i}^{n} w_i Y_i}_{=\mathrm{S}} \log\frac{\theta}{1-\theta} + \underbrace{\sum_{i}^{n} w_i}_{=\mathrm{N}} \log\,(1-\theta) \end{aligned} \tag{3.5}$$

$$\begin{aligned} \frac{dL(W,\theta)}{d\theta} &\stackrel{!}{=} 0 \\ \widetilde{\theta} &= \text{S/N} \,. \end{aligned} \tag{3.6}$$

With this at hand equation (3.3) can be reformulated to

$$\begin{aligned} \widetilde{f}(x) = \widetilde{\theta} &= \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{i=1}^{n} w_i(x) \log \mathbb{P}_\theta(Y_i) \\ &= \sum_{i}^{n} w_i Y_i \,. \end{aligned} \tag{3.7}$$

The obtained estimator (3.7) is in this case the Nadaraya-Watson estimator with uniform kernel weights w_i. Having estimated once a set of "weak" estimates $\{\widetilde{\theta}^{(k)}\}$ with $1 \leq k \leq K$ it is possible to estimate the unknown structure represented by the resulting estimates $\widehat{\theta}$. Whereas K denotes here the total number of obtained radii h. The underlying principle of the method assures that these last estimates perform at least as good as the best estimate of the above set of "weak" estimates $\{\widetilde{\theta}^{(k)}\}$.

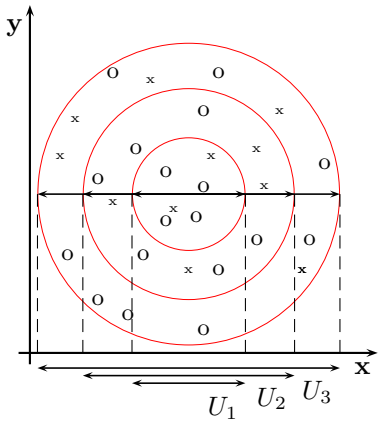

Fig. 3.2: Representation of three different window sizes on basis of uniform kernel functions. Each window size yields a new "weak" estimate $\widetilde{\theta}^{(k)}$ for the stagewise aggregation procedure.

Figure 3.2 illustrates the application of the uniform kernel weights to a cloud of data points. The present case shows the localization of $K = 3$ windows and the covered data points, which would yield an ordered sequence of local likelihood estimates $\{\widetilde{\theta}^{(k)}\}_{k=1,2,3}$ with decreasing variability (cf. equation (3.7)).

3.1.4.1 Technical Implementation Details

Implementation details concern the calculation of the "weak" estimates, where initially, distances $d(x, x')$ for all pairs of observations are computed. The obtained square distance matrix contains the distances between the n included objects. Distance measures for continuous data can be calculated by using the L_p-norms with $p \geq 1$

$$d(x, x') = \|x - x'\|_{L_p} = \left\{ \sum_{d=1}^{D} \|x_d - x'_d\|^p \right\}^{1/p}. \tag{3.8}$$

As it is common and sufficient for present purposes the L_2-norm (Euclidean norm with $p = 2$) is considered here for calculation of the distance matrix. At the very beginning of each calculation of distances with the L_p-norms, the included variables should be subject to a standardization with the objective of having all variables on the same scale. A suitable matrix for this issue may be $\mathcal{S} = \text{diag}(s_{x_1x_1}^{-1}, \ldots, s_{x_Dx_D}^{-1})$, the matrix of variances of all components. Altogether this yields

$$d(x, x') = \sum_{d=1}^{D} \frac{(x_d - x'_d)^2}{s_{x_dx_d}}. \tag{3.9}$$

For the localization process, the localizing schemes, in terms of the corresponding weights w_i, have to be selected. A convenient reformulation of equation (3.7) yields

$$\widetilde{\theta}_j^{(k)} = \frac{\sum_{i=1}^{n} y_i \mathbf{1}(\|x_j - x_i\| < h_j^{(k)})}{\underbrace{\sum_{i=1}^{n} \mathbf{1}(\|x_j - x_i\| < h_j^{(k)})}_{N_{h_j^{(k)}}}} .$$

Here $N_{h_j^{(k)}}$ denotes the local sample size, for example in the plane or ball of radius $h_j^{(k)}$. Similarly, $\widetilde{\theta}^{(k)}$ denotes the PD measured locally inside the ball of radius $h_j^{(k)}$. The Bernoulli random vector Y_i represents a known class assignment, in the used *Creditreform* data set, of solvent and insolvent firms. To select the neighborhood of points $\mathbf{1}(\|x_j - x_i\| < h_j^{(k)})$ that contributes to the estimation procedure at step j, an increasing sequence $h_j^{(1)} < \ldots < h_j^{(K)}$ of radii is generated. For instance the smallest ball of minimum radius $h^{(1)}$ defines here a set of observations containing at least a pair of two observations (solvent,insolvent) lying within this radius (cf. figure 3.3).

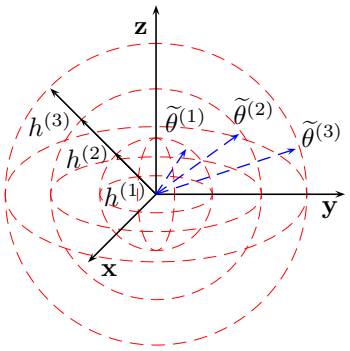

Fig. 3.3: Three dimensional representation of the generation of three "weak" estimates $\widetilde{\theta}^{(1)}, \ldots, \widetilde{\theta}^{(3)}$ from three balls of radii $h^{(1)}, \ldots, h^{(3)}$.

Belomestny and Spokoiny (2006) propose for calculation of this sequence to apply

another geometrically increasing sequence, i.e. for a r.v. $X_i \in I\!R^D$ it is

$$h^{(k)} = h^{(k-1)} \cdot \; a^{1/d} \quad \text{with} \quad 1.1 < a < 1.3 \quad (\text{here} \quad a = 1.25).$$

The generation of this sequence has to depend on the dimension d of the underlying r.v. X_i due to the *curse of dimensionality* problem, which says that observations in higher dimensions become more and more sparsely distributed so that even for large samples estimators based on local averaging perform more and more unsatisfactorily (cf. Härdle, Müller, Sperlich & Werwatz (2003)).

3.1.5 Stagewise Aggregation

With the calculated "weak" estimates $\widetilde{\theta}^{(k)}$, the SSA procedure has as its next crucial step the calculation of the final estimates $\widehat{\theta}$. The aim of the local adaptive estimation procedure is now to find an estimate $\widehat{\theta} = \widehat{\theta}(x)$ which performs as good as the best estimator of the sequence $\{\widetilde{\theta}^{(k)}\}$ of "weak" estimates. The central equation (3.10) here assembles in a recursive way at each iteration step k an aggregated or "improved" version of the resulting estimator $\widehat{\theta}^{(k)}(x)$ from the previous iteration step $k-1$. In other words, each new estimate $\widehat{\theta}^{(k)}(x)$ is a mix of the "weak" estimate $\widetilde{\theta}^{(k)}$ and the result $\widehat{\theta}^{(k-1)}$ of the foregoing iteration. Note here that at each iteration step k the radius h increases. The equation is

$$\widehat{\theta}^{(k)}(x) \;\; = \;\; \gamma_k \widetilde{\theta}^{(k)}(x) + (1-\gamma_k)\widehat{\theta}^{(k-1)}(x) \quad \text{with} \quad \widehat{\theta}^{(0)}(x) = \widetilde{\theta}^{(0)}. \qquad (3.10)$$

Equation (3.10) still contains several elements which need to be clarified in the following. Starting with the parameter γ_k which controls the degree to which the preceding $\widehat{\theta}^{(k-1)}(x)$ determines the current aggregate estimate $\widehat{\theta}^{(k)}(x)$. The parameter γ_k is intrinsically tied to both estimates of equation (3.10) and measures their

difference in terms of homogeneity of the local region. It is defined as

$$\begin{aligned} \gamma_k &= K_{tr}(\underbrace{m^{(k)}/\mathfrak{z}_k}_{=\nu}) \\ \text{with} \quad K_{tr}(\nu) &= (1-\nu)^+ \quad \text{and} \quad \nu > 0 \ . \end{aligned} \tag{3.11}$$

γ_k is the value from the triangle kernel K_{tr} that helps to decide whether the difference of both estimates is significant or not.

If it is significant, the new estimate $\widetilde{\theta}^{(k)}(x)$ is forced towards the previous estimate $\widehat{\theta}^{(k-1)}(x)$. Polzehl and Spokoiny (2005) called this action a "memory"-step which assures that the approximation bias stays on a moderate level when the neighborhood propagates.

3.1.5.1 The Kullback-Leibler Divergence

The parameter $m^{(k)}$ from equation (3.11), is the product of the Kullback-Leibler divergence (KLD), a distance value that measures the "distance" between two distributions P and Q, i.e. $\mathcal{K}(P, Q) = I\!E_P\{\log(dP/dQ)\}$, and the local sample size N_k. In terms of the parametric model $I\!P_\theta$ it is

$$\mathcal{K}(\widetilde{\theta}, \theta) = E_P\left\{\log\left(\frac{dP_{\widetilde{\theta}}}{dP_\theta}\right)\right\} \ . \tag{3.12}$$

From equation (3.5) together with $\mathrm{S} = \widetilde{\theta}\cdot\mathrm{N}$ it is possible to get the explicit expression of the fitted likelihood $L(W, \widetilde{\theta}, \theta) = L(W, \widetilde{\theta}) - L(W, \theta)$ with respect to the Bernoulli

law which enfolds the KLD

$$\begin{aligned} L(W,\widetilde{\theta}) - L(W,\theta) &= \widetilde{\theta} \cdot \mathrm{N} \cdot \log \frac{\widetilde{\theta}}{1-\widetilde{\theta}} + \mathrm{N} \cdot \log\,(1-\widetilde{\theta}) \\ &- \widetilde{\theta} \cdot \mathrm{N} \cdot \log \frac{\theta}{1-\theta} - \mathrm{N} \cdot \log\,(1-\theta) \qquad (3.13) \\ &= \mathrm{N}\left\{\widetilde{\theta} \cdot \log\left(\frac{\widetilde{\theta}}{\theta}\right) + (1-\widetilde{\theta}) \cdot \log\left(\frac{(1-\widetilde{\theta})}{(1-\theta)}\right)\right\} \\ &= N \cdot \mathcal{K}(\widetilde{\theta},\theta)\ . \qquad (3.14) \end{aligned}$$

Having obtained this result it is possible to assess the quality of the LPA (cf. Härdle and Spokoiny (2007)). With $\theta \in \Theta$ and $w_i > 0$ the modeling bias is defined as

$$\Delta(W,\theta) = \sum_{i=1}^{n} \mathcal{K}\{f(X_i), f_\theta(X_i)\}\mathbf{1}(w_i > 0). \qquad (3.15)$$

Then the so called "small modeling bias" (SMB) condition reads

$$\Delta(W,\theta) \leq \Delta\ . \qquad (3.16)$$

Now, in terms of the SMB it is possible at this stage to reformulate the aim of the approach: to discover the desired estimate $\widehat{\theta}$ and to select the "largest local" model with respect to weighting scheme and thus N_k, for which (3.16) still holds. The result is also known in this context as the "oracle"-estimator.

3.1.5.2 Generation of Critical Values $\mathfrak{z}_k$

To accomplish the presentation and to be able to run the procedure it is necessary to generate critical values $\mathfrak{z}_k$ that define the confidence set (CS) shown in figure 3.4.

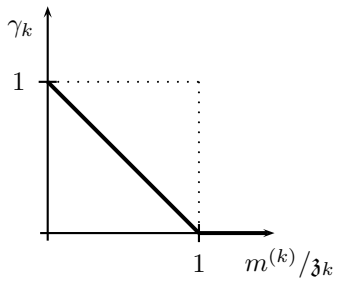

Fig. 3.4: Confidence area described by a triangle Kernel $K_{tr} = (1-\nu)^+$.

The CS is defined as

$$E_k(\mathfrak{z}) \quad = \quad \{\theta : N_k \mathcal{K}(\widetilde{\theta}^{(k)}, \theta) \leq \mathfrak{z}\}. \tag{3.17}$$

As long as (3.17) is fulfilled a stable region is existent and the estimator can propagate in the sense that no separation is done and $\widetilde{\theta}^{(k)}$ determines fully $\widehat{\theta}^{(k)}$. The calculation of these values has to be conducted with *each* variation of analysis variables or observations. The critical values $\mathfrak{z}_k$ now decide whether the KL-divergence is significant and violates the hypothesis of homogeneity, or not. In this case the current estimate $\widehat{\theta}^{(k)}$ is the closest to the aforementioned "oracle"-estimate $\widetilde{\theta}^{(k)}$ holding exactly the LPA. Practically, a sequence of values $\{\mathfrak{z}_k\}$ can be obtained by resampling. Belomestny and Spokoiny (2006) propose a simplified parameter choice approach for the generation of these values and give, moreover, accurate recommendations concerning the calibration of the parameter. Following their instructions the determination of the values should be started with the least value $\{\mathfrak{z}_K\}$ of the sequence and its corresponding "weak" estimates $\widetilde{\theta}^{(K)}$ and $\widetilde{\theta}^{(K-1)}$, so, that alternatively $\widehat{\theta}(\mathfrak{z}_K)) = \gamma\widetilde{\theta}^{(K)} + (1-\gamma)\widetilde{\theta}^{(K-1)}$ can be computed, with $\gamma = K_{tr}(m/\mathfrak{z}_K)$ and $m = N_K \mathcal{K}(\widetilde{\theta}^{(K)}, \widetilde{\theta}^{(K-1)})$. Furthermore

$$\sup_{\theta^* \in \Theta} I\!E_{\theta^*} |N_K \mathcal{K}(\widetilde{\theta}^K, \widehat{\theta}(\mathfrak{z}_K))|^r \quad \leq \quad \rho \mathfrak{r}_r/(K-1) \tag{3.18}$$

should hold for a choice of $\mathfrak{r}_r = 2r \int_{\mathfrak{z}\geq 0} \mathfrak{z}^{r-1} e^{\mathfrak{z}} d\mathfrak{z} = 2r\Gamma(r)$, $\rho = 1$, $r = 1/2$ and $\theta^* = 0.5$. If an accurate value of $\mathfrak{z}_K$ is found the other $\mathfrak{z}_k$ values can be found in the form $\mathfrak{z}_k = \mathfrak{z}_K + \iota(K-k)$. Now, a complete set of values is obtained which needs to fulfill additionally

$$\sup_{\theta^* \in \Theta} I\!E_{\theta^*} |N_k \mathcal{K}(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k)})|^r \quad \leq \quad \rho \mathfrak{r}_r \ . \tag{3.19}$$

Theoretical foundations and a further detailed treatise of this approach can be found in Belomestny and Spokoiny (2006).

3.1.6 Description of the Algorithm and Summary

Together with a visualization of the SSA algorithm, the above stated mathematical results should be summarized to show a realization of the method.

Module 1: Choose $Z_i = (Y_i, X_i)$ with $i = 1, \ldots, n$ - two vectors of random variables, where X_i reflects the explanatory variables or locations, valued in the finite Euclidean space $\mathcal{X} = \mathbb{R}^D$ and the Y_i's as responses, conditioned on $X_i = x$.

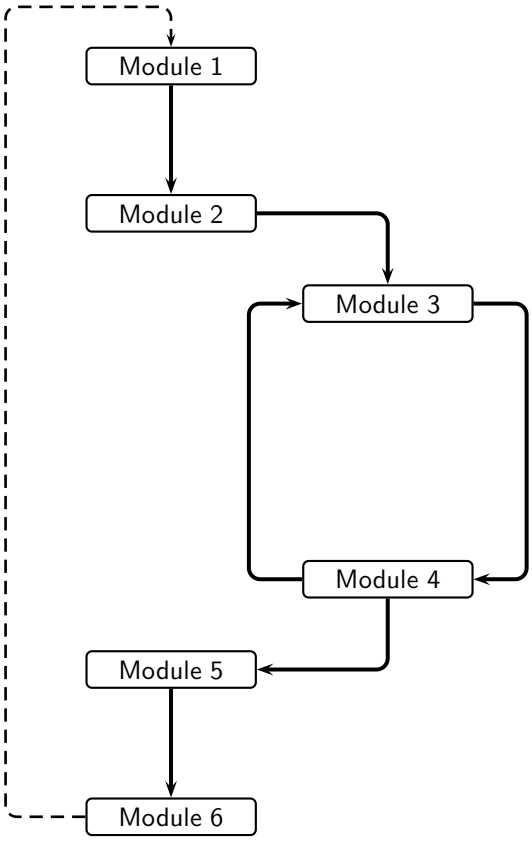

Fig. 3.5: Module structure of the implemented SSA algorithm.

Module 2: Generate distances $d(x, x')$ for all pairs of observations.

Module 3: Calculate $\{\widetilde{\theta}^{(k)}\}$, $k = 1, \ldots, K$, a sequence of weak local likelihood estimates at a point x.

Module 4: Perform Monte-Carlo simulations of the "weak" estimates to obtain a sequence of "critical"-values $\{\mathfrak{z}_k\}$, $k = 1, \ldots, K$.

Module 5: Stagewise aggregation process is started: initialization with $\widehat{\theta}^{(1)}(x) = \widetilde{\theta}^{(1)}(x)$. An aggregate estimate $\widehat{\theta}$ is constructed by equation (3.10).

Module 6: Evaluation of results and generation of graphical output.

Loop: If desired repeat procedure several times for alternative evaluations of different data sets.

This modular structure approximately reflects the implemented SSA algorithm in the Statistical Analysis System (SAS).

3.2 Support Vector Machines

The support vector machine (SVM) approach comes from the field of machine learning and one of the first advances regarding this method was done by Tikhonov and Arsenin (1977) and later by Vapnik (1995). The motivation of doing research in this domain was to increase the accuracy of estimation results of classification methods. In this regard SVM appeared to have promising properties as it is able to select a classifying function based on very general criteria. Moreover its solution is unique and flexible and is controlled only by few parameters (cf. Härdle, Moro & Schäfer (2004)). Results from this method will be delivered as a comparison to the SSA approach in section 5. The underlying idea of SVM will be outlined in the following according to Moro (2004) who gives for further reading a detailed treatise on this topic.

3.2.1 Theoretical Aspects of SVM

The theoretical idea of the SVM wil be presented for a linearly non-separable case as this case corresponds most to the situation encountered in praxis. At the core of the SVM approach lies a classification function of the form

$$f \quad = \quad \left\{ x^\top \omega + b; \quad \text{margin} \rightarrow \max_{\omega, b}, \quad \nexists i : x_i \in \text{margin zone} \right\} . \qquad (3.20)$$

The crucial element of SVM is that it tries to find a certain separating hyperplane $x^\top \omega + b$ that offers the largest possible margin of two observations in two distinct classes, for instance solvent and insolvent companies. Figure 3.6 illustrates in a striking way this separation process.

Here the canonical hyperplanes $x^\top \omega + b = \pm 1$ define a line above and below the separating hyperplanes and are boundaries of the corresponding classes ($y = +1$

and $y = -1$). The distance between those hyperplane is called the margin, which equals $2/||\omega||$. Moreover, $||\omega||$ defines the Euclidean norm. As this is the linearly non-separable case, the slack variable $\xi \geq 0$ defines a misclassification error as the distance to the boundary of the class to which the observation belongs ($\xi/||\omega||$). The parameter ξ is positive in case of a misclassified observation and 0 otherwise.

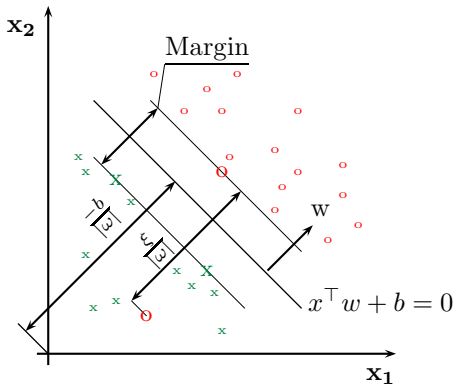

Fig. 3.6: The separating hyperplane $x^\top \omega + b = 0$ and the margin in the linearly non-separable case (Source: Moro(2004)).

While solving the convex optimization problem the following conditions must be satisfied for all $i = 1, 2, \ldots, n$:

$$y_i(x^\top \omega + b) \geq 1 - \xi_i \tag{3.21}$$

$$\xi_i \geq 0 \; . \tag{3.22}$$

It is worth noting that the presented SVM framework can be generalized to the nonlinear case via the application of kernels. In the nonlinear case it would be possible to discriminate more precisely than in the linear case between observations from two separated classes. This is exactly what was done for the generation of the results presented in section 5.

The dual SVM optimization problem in the non-linear case uses a kernel function $K(x_i, x_j)$ and is based on the dual Lagrangian:

$$L_D \;=\; \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \; \alpha_j \; y_i \; y_j \; K(x_i, x_j) \; . \tag{3.23}$$

Here alpha are Lagrange multipliers. The dual problem is:

$$\max_{\alpha_i} L_D,$$

subject to

$$0 \leq \alpha_i \leq C,$$
$$\sum_{i=1}^{n} \alpha_i \; y_i = 0 \; .$$

Possible kernel Gaussian functions in this context would be for instance an isotropic Gaussian Kernel or a stationary Gaussian Kernel. It should be mentioned here that parameters for the SVM evaluation were not optimized but suggested as in Chen, Härdle & Moro (2006), which yields for the radial basis $RB = 2.0\Sigma^{1/2}$ and for the complexity parameter $C = 1.0$.

4 Data and Variable Selection

The applied *Creditreform* dataset was obtained from the of the *Financial and Economic Data Center* (FEDC) database. It contains 21000 randomly selected observations of German firms from different branches and covers the years $1996 - 2002$. Each record provides information concerning the firm's economic situation, the size, in terms of sales and employment levels, and the legal form of a company. The data set is divided in two parts: the *first* part contains 1000 observations of firms two years prior to their insolvency. The *second* part enfolds the remaining 20000 records of solvent firms.

4.1 Data Selection and Preprocessing

The purpose of this work is to derive statements concerning the expected costs of bankruptcy and the optimal capital structure of a firm. However, the expected costs of bankruptcy may differ with respect to the branch of a firm. From a theoretical point of view this is solely due to the fact, that the expected costs of bankruptcy were defined as a function of the loss given default (LGD), which describes the percentage of exposure at default (EAD) that will not be recovered following insolvency of a firm. This percentage will be higher as a firm of the service sector becomes bankrupt, compared to that of a firm from the manufacturing industry sector. Hence, it is necessary to analyze different sectors separately.

For the application of classification methods and the generation of default probabilities it is necessary to possess the maximum information in terms of observations, above all from firms that went bankrupt. The number of 1000 in the present case represents the first restriction as the classification procedures, SSA as well as SVM, demand the selection of a subset of records for the *training data* set and the *validation data* set. The random selection of both data sets is done without replacement

and with each class (solvent or insolvent) representing 50% of the data set. This fact is a challenge for the construction of an accurate design of the analysis framework. The following approach is motivated by this fact: to extract the maximum information it is convenient to choose firms from the manufacturing industries (approximately 25% of all observations) which, to a greater or lesser extent, represent a complete set of homogeneous branches. Furthermore this solution accommodates with the initial motive: to obtain information about the optimal capital structure of a firm. Accordingly this information can be derived in the following for German manufacturing firms. To the German manufacturing industries the numbers $15-36$ are assigned as the last two digits of the WZ 93 classification code. The companies so classified will be selected for analysis. Even though this group constitutes the biggest possible of the whole data set, only approximately 300 observations of insolvent firms are left, which is not very much for the generation of *training* and *validation* data sets (preferably $400-500$ observations). Unfortunately this requires the use of all insolvent firms in the subsamples. Solvent companies do not pose such limitations since they are much more numerous.

On this subset of records the following preprocessing was performed: extreme values, outliers were replaced by the 5% (resp. 95%) percent quantile. For the generation of the financial ratios which will be mainly for the analysis, those observations had to be removed which had missing values or appeared in the denominator and had a zero value. Moreover observations of solvent firms from the year 1996 had to be removed as counterpart values of insolvent firms were not available. After data preprocessing the data set had 714 observations of insolvent firms and 4392 observations of solvent firms.

4.2 Financial Ratios

The *Creditreform* bankruptcy data set contains sufficient information for the generation of financial ratios. According to Chen, Härdle & Moro (2006), whose work is likewise based on the *Creditreform* bankruptcy data set and presents the performance of the SVM classification analysis, six categories of accounting ratios can be distinguished: ratios concerning the leverage, profitability, liquidity and activity of a firm, firm size and the percentage of change for some variables. Furthermore their work offers an additional survey concerning the profile of the data. An important upcoming issue is to detect the most promising predictors for the SSA classification method. As it was pointed out in the previous section the SSA method relies fundamentally on nonparametric theory and is consequently subject to the *curse of dimensionality* (cf. page 17). Therefore the number of variables used for classification should not be too extensive. An adequate choice of variables would be those variables that were highlighted as the most promising predictors in Chen, Härdle & Moro (2006).

Abbreviation	Variables
EBIT	Earnings before interests and taxes
TL	Total liabilities
TA	Total assets
IDINV	Increase (decrease) of the inventories
AP	Accounts payable
INV	Inventories
CASH	Cash and cash equivalents
OI	Operating income
NI	Net income
SALES	Total sales

Tab. 4.1: Selection of variables for ratio calculation (Source: Chen et al. (2006)).

Table 4.1 presents those variables delivered from the *Creditreform* data set which were employed to calculate the eight most promising financial ratios reported in table 4.2 together with the corresponding category.

Variables	Ratio	Class
1	EBIT/TA	Profitability
2	TL/TA	Leverage
3	IDINV/INV	Percentage of Incremental Inventories
4	AP/SALES	Account Payable Turnover Activity
5	INV/SALES	Inventory Turnover Activity
6	CASH/TA	Liquidity
7	OI/TA	Profitability
8	NI/SALES	Net Profit Margin Profitability

Tab. 4.2: Utilized variables for the backward selection process.

5 Empirical Results

5.1 Variable Selection Results

Variable selection results are presented in table 5.1 as well as in figure 5.1 in form of a box-plot. The backward selection procedure, as it is understood here, is the process where for each set of 100 iterations one variable is withdrawn (starting with all listed variables in table 4.2 and ending with the two remaining ones). As reported in table 5.1 the procedure selects four financial ratios with the highest values for median (69%) and overall mean (67.5%) of all calculated hit rates or the percentage of correctly classified companies. The distribution of the HR values is well illustrated in figure 5.1 with the Box-Plots. More precisely, the employed ratios are listed in table 4.2, where the column *variables* indicates the number of employed ratios.

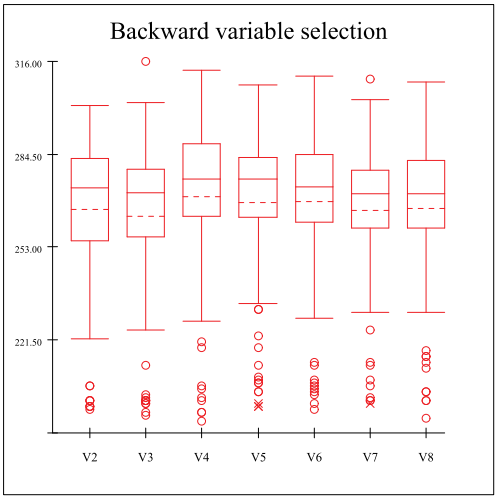

Fig. 5.1: Backward selection starting with the maximum of eight variables ($d = 8$) denoted as V8 and ending with a bivariate model ($d = 2$) denoted as V2. The number of replications is 100, the selection criterion at each step was the highest median hit rate. Dashed lines define the mean, solid lines the median values as reported in table 5.1.

Additionally table 5.1 summarizes all obtained results from the backward selection process, which are illustrated in figure 5.1, in a detailed way.

Variables	Median (%)	Mean (%)	Min (%)	Max (%)
2	273.0 (68.3)	265.7 (66.4)	205 (50.1)	301 (75.3)
3	271.5 (67.9)	263.6 (65.9)	204 (50.1)	316 (79.0)
4	276.0 (69.0)	270.2 (67.5)	210 (50.2)	313 (78.3)
5	276.0 (69.0)	268.1 (67.0)	209 (78.3)	308 (77.0)
6	273.5 (68.4)	268.3 (67.1)	213 (53.3)	311 (77.8)
7	271.0 (67.8)	265.6 (66.4)	212 (53.0)	310 (77.5)
8	271.0 (67.8)	266.3 (66.6)	217 (54.3)	309 (77.3)

Tab. 5.1: Summary hit rate values from backward selection procedure.

Table 5.2 reports additionally the most important location parameter for the four chosen variables, which will be employed for classification.

Variable	EBIT/TA	TL/TA	AP/SALES	IDINV/INV
Minimum	-0.097	0.000	0.012	-0.492
First quantile	0.015	0.235	0.034	-0.046
Median	0.054	0.499	0.060	0.000
Mean	0.068	0.475	0.079	0.018
Third quantile	0.113	0.740	0.106	0.098
Maximum	0.290	0.915	0.236	0.499

Tab. 5.2: Location parameter for the selected variables.

5.2 Comparison of the SSA and SVM Results

Two important categories represented by the selected variables in table 4.2 are profitability and leverage. Both types deliver highly relevant information concerning the actual situation of a firm. For instance, solvent firms tend to have a positive profitability, in terms of EBIT per total assets, and a reasonable leverage, in terms of total liabilities per total assets, of about 0.4 and more (cf. figure 5.2 and table 5.2).

SSA and SVM classification results are presented as two dimensional colored plots of the default probabilities. Stronger red colors indicate regions of insolvent firms (white circles), whereas green colored regions stand for solvent firms (black triangles). Classification results were obtained on the basis of a combination of four variables (cf. table 4.2), therefore the two dimensional color plots represent hit rates only for a two dimensional plane with two variables fixed. Nevertheless both SSA illustrations 5.2 and 5.4 show strong tendencies of a correct classification, whereas it identifies on a 50% level (blue line) in a more reliable way insolvent firms than solvent firms. However, a clearer discrimination of insolvent and solvent firms is done on a 70% (resp. 65%) probability of default level (white line). Figures 5.2 and 5.3 illustrate on a two dimensional basis the classification results of SSA and SVM. As the variable choice concerns a total of four ratios the remaining two variables were replaced by their median values. Both figures plot profitability and leverage with the same set of observations. White circles show the solvent firms and black triangles the insolvent ones. Moreover for a better intuition two separation lines 50% (blue) and 70% (white) probability of default are plotted. At first glance the SVM illustration defines the separation areas more cleanly in terms of colors, i.e. default probabilities and classification. Where the SSA classification result (figure 5.2) suggests a more or less definite class separation by the white line, the SVM

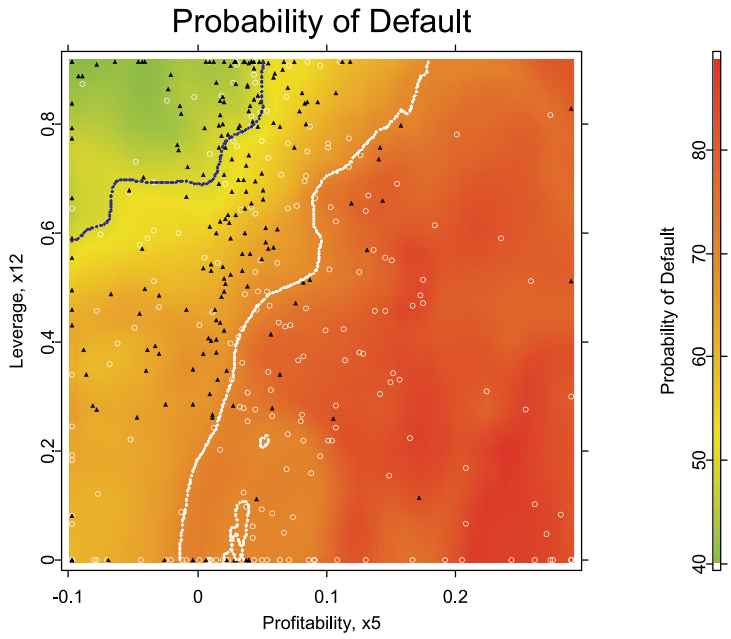

Fig. 5.2: Probability of default estimated for a random subsample of 400 insolvent and 400 solvent companies. The plot presents the variables $x5$ and $x12$ and the blue 50% (resp. white 70%) separation line. The estimation procedure was SSA and classification HR = 77%.

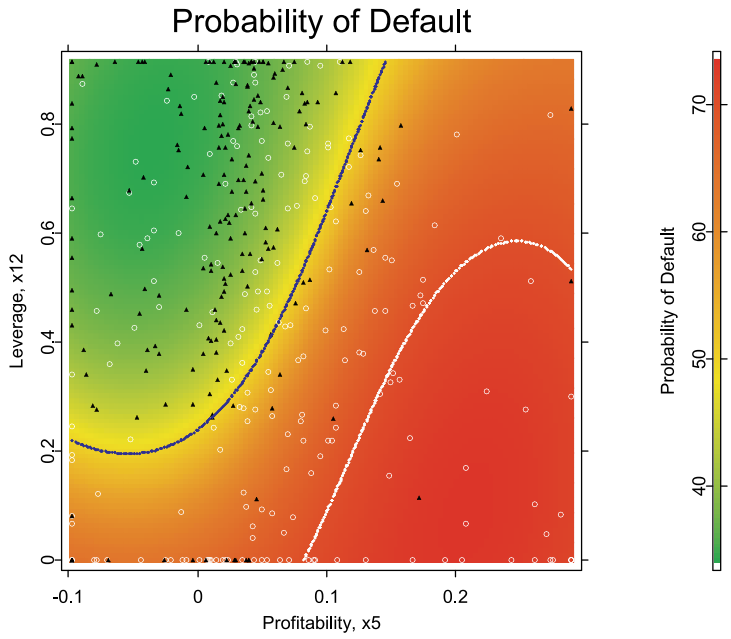

Fig. 5.3: Probability of default estimated for a random subsample of 400 insolvent and 400 solvent companies. The plot presents the variables $x5$ and $x12$ and the blue 50% (resp. white 70%) separation line. The estimation procedure was SVM ($RB = 2.0\Sigma^{1/2}, C = 1.0$) and classification HR = 80%.

does this in a more intuitive way by the blue line. Recall that the 50% line (blue) represents the boundary between solvent and insolvent companies whereas the 70% line (white) is reported for comparison purposes.

Figures 5.4 and 5.5 illustrate again on a two dimensional basis the classification results of SSA and SVM. Yet, here profitability is plotted against the account payable turnover. Here the SSA method defines a larger region for the solvent firms, but in comparison with the SVM again a slightly different configuration appears, as SVM suggests to discriminating in a more horizontal way. On closer inspection this separation turns out to be more intuitive, as on the upper right part of the picture more solvent firms are located.

	SVM		SSA	
Variable	Hit Rate	(%)	Hit Rate	(%)
Minimum	285.0	(71.3)	210.0	(50.2)
Median	308.5	(77.1)	276.0	(69.0)
Mean	308.4	(77.1)	270.2	(67.5)
Maximum	324.0	(81.0)	313.0	(78.3)

Tab. 5.3: Hit rate summary values from SVM and SSA classification methods for 100 randomly selected validation subsets of 400 observations.

Table 5.3 summarizes the SVM classification results together with those for the SSA for 100 randomly selected validation subsets of 400 observations from the *Creditreform* data. In comparison to the SSA results of table 5.1 the indicated suspicion of a superior performance of SVM in terms of hit rates is affirmed. A noticeable difference appears not only as the median is 8% higher (equivalent to a difference of 32 correct classified firms) but also with the fact that median and mean values are virtually identical indicating the symmetry of the distribution.

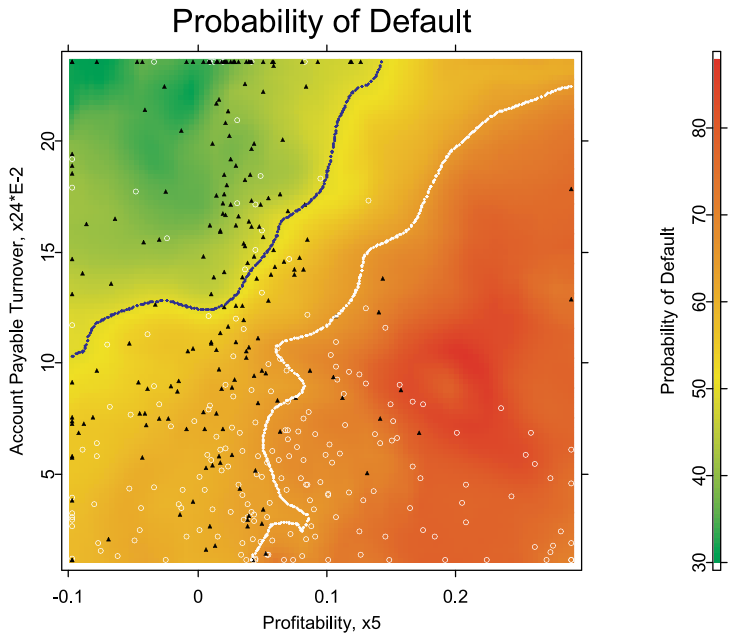

Fig. 5.4: Probability of default estimated for a random subsample of 400 insolvent and 400 solvent companies. The plot presents the variables $x5$ and $x24$ and the blue 50% (resp. white 65%) separation line. The estimation procedure was SSA and classification HR = 77%.

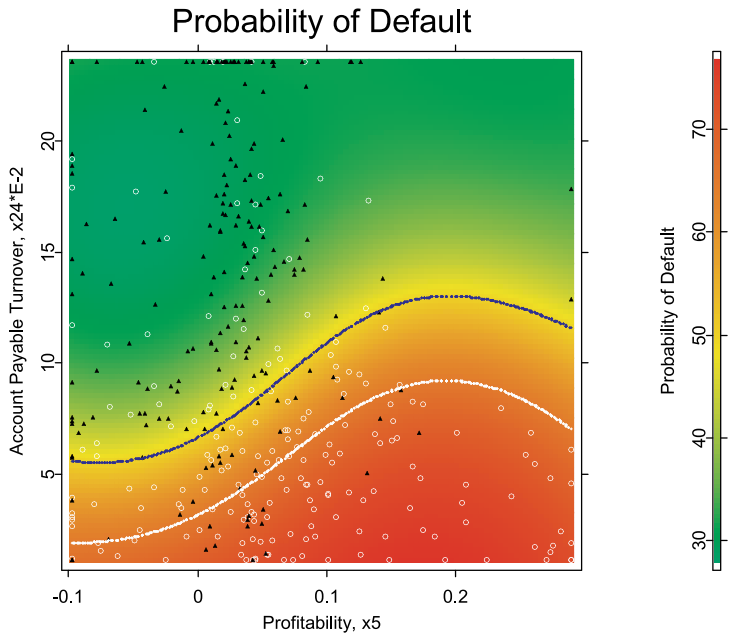

Fig. 5.5: Probability of default estimated for a random subsample of 400 insolvent and 400 solvent companies. The plot presents the variables $x5$ and $x24$ and the blue 50% (resp. white 65%) separation line. The estimation procedure was SVM ($RB = 2.0\Sigma^{1/2}, C = 1.0$) and classification HR = 80%.

Irrespective of the hit rate performance of both methods the graphical representation suggests a higher quality of the SVM results. This is due on the one hand to the fact that separation lines appear to be more intuitive and on the other hand to the fact that the shape of the separation lines are much less complex than those of the SSA method. An explanation might be that the SVM can optimize the complexity of the classifying function better than the SSA which possibly discards the hypothesis of homogeneity too often. In other words the significance level of the homogeneity tests in the SSA may be too low (too many rejections of H_0). The evidence of that comes precisely from the irregular boundary between the classes that was produced by the SSA. The one computed by the SVM is much smoother.

Another reason may be that the SVM looks at the proximity between observations in terms of the score or PD. Thus, the Euclidian distance $||x - x'||$ between two observations can be large and yet these two observations be close in terms of PD. The SSA computes PDs based on Euclidian distances only.

It should be mentioned here that the performance of the SVM may still be increased by adjusting its parameters: the radial basis coefficient and capacity.

5.3 WACC Minimization Problem

The optimal capital structure formulated in section 2.1 as a minimization problem and therewith the estimation of the theoretical relationships presented in figure 2.2, will be discussed in this part. Necessary elements for the solution were $(1 - \mathrm{TS}/TA)$ and equation (2.4), which could be calculated on the basis of the *Creditreform* data and the estimated default probabilities from the SSA method on basis of correct classifications of a randomly selected subsample of 400 observations from the manufacturing industries.

To obtain a real minimum solution it is necessary that at least one of the curves is estimated nonlinearly. This was done by considering elementary relationships, i.e.

$$\begin{aligned} & \mathbf{f_1} &=& \; a_1 + b_1 \cdot x + c_1 \cdot x^2 \\ \text{and} \quad & \mathbf{f_2} &=& \; a_2 + b_2 \cdot x \; . \end{aligned}$$

The minimum of the third function $f_3 = f_1 + f_2$ could be obtained first analytically and then in an empirical way.

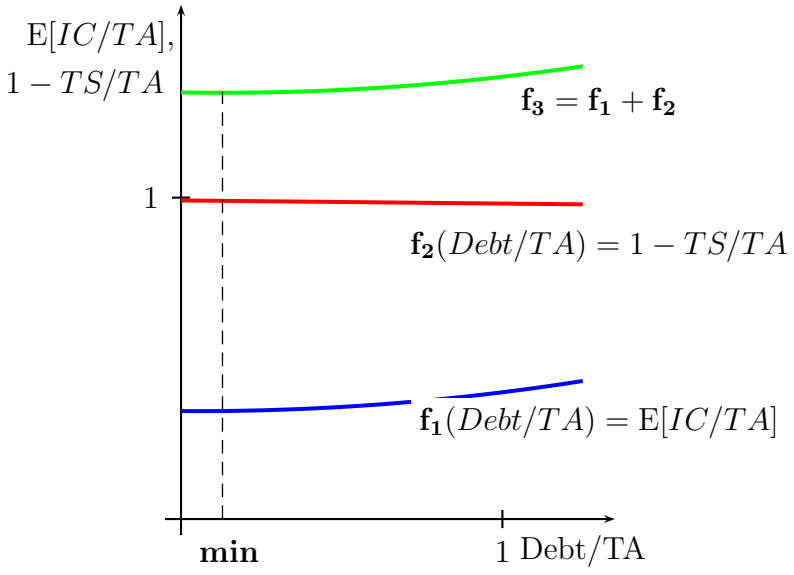

Fig. 5.6: Summary plot of the estimated functions f_1, f_2 and f_3.

Figure 5.6 illustrates the estimated curves. The estimation of $\mathbf{f_2}$ appears to have a nearly constant value, though the slight trend turns out to be correct with respect to the fact that the y-axis presents $1-TS/TA$, which is essentially, up to a constant, forgone savings from the tax shield. The tax shield is varying between zero and five percent of total assets. The underlying relationship is supposed to be linear as the tax-shield increases with higher leverage, thus justifying the linearity of $\mathbf{f_2}$.

Coefficient	$\mathbf{f_1}$	$\mathbf{f_2}$
a	0.3360	0.9914
b	-0.0069	-0.0098
c	0.0655	

Tab. 5.4: Minimization problem estimation results.

Together with the results from table 5.4 the minimum is

$$\left(\frac{DEBT}{TA}\right)^* = -\frac{b_1+b_2}{2c_1} = 13\%. \tag{5.1}$$

This means, that on the basis of the estimated data the debt fraction of 13% of total assets of a firm in the German manufacturing industries would yield the optimal firm value. This value is lower than the calculated median value of $DEBT/TA$ which was about 18.8%. Compared to values found in previous publications, for instance Ju et al. (2005) who predicted on the basis of US data the value of 15.3% the obtained 13% for German firms appear to be reasonable. According to Ju et al. (2005) the discrepancy between this value and the median can be interpreted as a too high debt financing that emerges as many small and medium sized businesses use the external financing possibility too extensively. Furthermore, it should be mentioned in this context that comparison calculations on the basis of the PDs estimated with the SVM did not change the obtained result in a significant way.

6 Conclusion

This thesis introduces a recently developed classification method namely the Spatial Stagewise Aggregation, which is based on nonparametric theory. Additionally for performance comparing reasons another well performing classification method, the Support Vector Machine, was presented. The performance comparison on the basis of the hit rate, as the percentage of correctly classified observations, showed a clear domination of the SVM method. Moreover, simulations showed a nearly symmetric distribution of SVM classification hit rates and a smaller range, i.e. the reliability turns out to be higher for SVM.

With respect to the formulated capital structure problem, moreover, a reasonable value of 13% of debt financing could be obtained, which was in line with results from other publications.

References

Belomestny, D. and V. Spokoiny, 2006: Spatial aggregation of local likelihood estimates with applications to classification. *SFB 649 Discussion Paper*, **036**.

Breiman, L., 1996: Bagging predictors. *Machine Learning*, **24**, 123–140.

Breiman, L., 1998: Arcing classifiers. *Annals of Statistics*, **26**, 801–849.

Chen, S., W. K. Härdle, and R. A. Moro, 2006: Calculation of default probabilities with support vector machines. *SFB Discussion Paper*.

Damodaran, A., 2002: *Investment Valuation*. John Wiley & Sons, New York.

Fan, J., M. Farmen, and I. Gijbels, 1998: Local maximum likelihood estimation and inference.

Freund, Y., 1995: Boosting a weak learning algorithm by majority. *Information and Computation*, **121**, 256–285.

Härdle, W., 1990: *Applied Nonparametric Regression*. Springer Verlag Heidelberg.

Härdle, W., M. Müller, S. Sperlich, and A. Werwatz, 2003: *Nonparametric and Semiparametric Models*. Springer Verlag Heidelberg.

Härdle, W., R. A. Moro, and D. Schäfer, 2004: *Statistical Tools for Finance and Insurance*. Springer Verlag, Berlin.

Härdle, W., R. A. Moro, and D. Schäfer, 2007: Estimating probabilities of default with support vector machines. *The Journal of Banking and Finance*.

Härdle, W. and V. Spokoiny, 2007: Foundations and applications of modern nonparametric statistics.

Ju, N., R. Parrino, A. M. Poteshman, and M. M. Weisbach, 2005: Horses and rabbits? trade-off theory and optimal capital structure. *Journal of Financial and Quantitative Analysis*, **40**, 259–281.

Moro, R. A., 2004: Rating companies with support vector machines. Master's thesis, Humboldt-University Berlin.

Polzehl, J. and V. Spokoiny, 2000: Adaptive weights smoothing with applications to image restoration. Sonderforschungsbereich 373 1998-77, Humboldt Universität zu Berlin. available at http://ideas.repec.org/p/wop/humbsf/1998-77.html.

Polzehl, J. and V. Spokoiny, 2005: Propagation-separation approach for local likelihood estimation. *Probability Theory Relatede Fields* DOI:10.1007/s00440–005–0464–1.

Ross, S. A., R. W. Westerfield, and B. D. Jordan, 2003: *Fundamentals of Corporate Finance alternate Edition*. McGraw-Hill/Irwin.

Tikhonov, A. and V. Arsenin, 1977: *Solution of Ill-posed Problems*. W.H. Winston.

Vapnik, V. N., 1995: *The Nature of Statistical Learning Theory*. Springer.

Ehrenwörtliche Erklärung

Ich versichere hiermit, dass ich die vorliegende Masterarbeit mit dem Thema:

"*Analysis of the Capital Structure of German Companies with the SSA and SVM*"

selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen wurden, habe ich in jedem einzelnen Fall durch die Angabe der Quelle, auch der benutzten Sekundärliteratur, als Entlehnung kenntlich gemacht.

Ort, Datum Unterschrift

Berlin,